Cloud Data Warehouse Performance Testing

Prepared by:
McKnight Consulting Group
www.mcknightcg.com
January 2021

Contents

Executive Summary

Big data analytics platforms load, store, and analyze volumes of data at high speed, providing timely insights to businesses. Data-driven organizations leverage this data, for example, for advanced analysis to market new promotions, operational analytics to drive efficiency, or for predictive analytics to evaluate credit risk and detect fraud. Customers are leveraging a mix of relational analytical databases and data warehouses to gain analytic insights.

This report focuses on relational analytical databases in the public cloud because deployments are at an all-time high and poised to expand dramatically. The cloud enables enterprises to differentiate and innovate with these database systems at a much more rapid pace than was ever possible before. The cloud is a disruptive technology, offering elastic scalability vis-à-vis on-premises deployments, enabling faster server deployment and application development, and allowing less costly storage. For these reasons and others, many companies have leveraged the cloud to maintain or gain momentum as a company.

This report outlines the results from an analytic performance test derived from the industry standard TPC Benchmark™ DS (TPC-DS)[1] to compare Cloudera Data Warehousing service (or CDW), which is part of the broader Cloudera Data Platform, Amazon Redshift, Azure Synapse Analytics, and two anonymized, public cloud-based data warehouses. Overall, the test results were insightful in revealing query execution performance of these platforms.

In terms of price per performance, Cloudera ran the Field Test 20% cheaper than the nearest competitor, Amazon Redshift, 40% cheaper than Synapse, 80% cheaper than one of the anonymized competitors and 5.5 times cheaper than the other anonymized competitor.

[1] More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

Introduction

Performance is important but is only one criterion for a data warehouse platform selection. This is only one point-in-time check into specific performance. There are numerous other factors to consider in selection across factors of Administration, Integration, Workload Management, User Interface, Scalability, Vendor, Reliability, and numerous other criteria. It is also our experience that performance changes over time and is competitively different for different workloads. Also, a performance leader can hit up against the point of diminishing returns and viable contenders can quickly close the gap.

MCG runs all of its performance tests to strict ethical standards. The results of the report are the objective results of the application of queries to the simulations described in the report. The report clearly defines the selected criteria and process used to establish the field test. The report also clearly states the data set sizes, the platforms, the queries, etc. used. The reader can determine how to qualify the information for their individual needs. The report does not make any claim regarding third-party certification and presents the objective results received from the application of the process to the criteria as described in the report. The report strictly measures performance and does not purport to evaluate other factors that potential customers should find relevant when making a purchase decision.

This is a sponsored report. Cloudera chose the competitors, the test and the Cloudera cluster size. The default configurations were chosen. MCG set up the environments, and ran the queries. Choosing compatible configurations is subject to judgment. We have attempted to describe our decisions in this paper.

In this writeup, all the information necessary is included to replicate this test. You are encouraged to compile your own representative queries, data sets, data sizes and compatible configurations and test for yourself.

Cloud Analytics Platform Offerings

This report outlines the results from an analytic performance test derived from the industry standard TPC Benchmark™ DS (TPC-DS)[2] to compare Cloudera Data Warehouse (running Impala-based virtual warehouses), Amazon Redshift, Azure Synapse Analytics, and 2 anonymized data warehouses—five relational analytical databases based on scale-out cloud data warehouses and columnar-based database architectures. Despite these similarities, there are some distinct differences in the platforms.

Cloudera Data Platform

The only single platform offering we tested that boasts flexibility through support for both data center and multiple public cloud deployments as well as capabilities across analytical, operational, data lake, data science, security, and governance needs is Cloudera Data Platform (CDP).

CDP is a secure and governed cloud service platform that offers a broad set of enterprise data cloud services with the key data functionality for the modern enterprise. CDP was designed to address multi-faceted needs by offering multi-function data management and analytics to solve an enterprise's most pressing data and analytic challenges in a streamlined fashion.

The architecture and deployment of CDP begins with the Management Console. The Management Console is where several important tasks are performed. First, the preferred cloud environment, e.g., AWS or Azure, is set up. Second, Data Warehouse clusters, and Machine Learning workspaces are launched. Third, additional services, such as, Data Catalog, Workload Experience Manager, and Replication Manager are utilized if required.

The Cloudera Data Warehouse service provides self-service independent virtual warehouses running on top of the data kept in a cloud object store, such as S3. The virtual warehouses use Data Catalog as a logical collection of metadata to define the managed data and its business context. Multiple virtual warehouses can share a single Data Catalog. The advantages of this virtual warehouse architecture include isolation and automatic configuration. Virtual warehouses and their compute resources are isolated to prevent "noisy neighbors" or resource hog queries bogging down a conventional monolithic data warehouse. Virtual warehouses also have automated management and performance scaling features, such as auto-scaling, auto-suspend, and auto-resume. These features simplify capacity planning, adjust compute capacity

[2] More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

to workload requirements, and ensure that you only pay for what you need to run your queries.

Amazon Redshift

Amazon Web Services Redshift was the first managed data warehouse cloud service and continues to get a high level of mindshare in this category. It indeed ticks all the table stakes boxes for a cloud analytic database. Amazon Redshift is a fit for organizations needing a data warehouse with little to no administrative overhead and a clear, consistent pricing model. Redshift (when run without Spectrum) is different from all the other competitors in that it doesn't read directly from cloud object storage during query processing (a separate loading step is required).

We tested the latest RA3 Redshift engine which introduced a new managed storage layer, which is an upgrade from the tighter coupled storage on the older DS2 and DC2 instance types. For Redshift, we paid an hourly rate for when the cluster was running, but it also has a pause feature to stop billing. Even with managed storage Redshift is still different from the other competitors in that it doesn't read directly from object storage (e.g. S3) during query processing and that data must be explicitly loaded from S3 into the managed storage and the time required for that loading wasn't included in the benchmark.

Azure Synapse Analytics

On Azure Synapse Analytics, formerly known as Azure SQL Data Warehouse, storage is separate from the compute Data Warehouse Unit (DWU). This enables Azure Synapse to scale columnar storage capacity and compute resources independently. This capability adjusts to various workload demands, offering potential cost savings when demand is low. Synapse can pause and resume compute billing, where only storage is billed during the paused time. Synapse achieves good balance in both configurability and simplicity, in a way that is both easy to administer and flexible in handling almost any usage pattern.

With Synapse, you can scale the compute DWU on the fly. We paid an hourly rate for when our cluster was active, but there is also a separate data storage charge for the SQL database underneath the Synapse engine.

Test Setup

The data sets used in the test were a workload derived from the well-recognized industry standard TPC Benchmark™ DS (TPC-DS).

From tpc.org: "The TPCDS is a decision support benchmark that models several generally applicable aspects of a decision support system, including queries and data maintenance. The benchmark provides a representative evaluation of performance as a general-purpose decision support system… The purpose of TPC benchmarks is to provide relevant, objective performance data to industry users. TPC-DS Version 2 enables emerging technologies, such as Big Data systems, to execute the benchmark."

The parameter values for the queries used across all vendors are from the TPC Benchmark™ DS (TPC-DS)[3] 2.13 spec validation queries. **This is not an official TPC benchmark.** The queries were executed using the following setup, environment, standards, and configurations.

The data model consists of 24 tables—7 fact tables and 17 dimensions. To give an idea of the data volumes used in our field test, the following table gives row counts of fact tables in the database when loaded with 30TB of data:

Table 1. Database Row Count

TPC-DS Table	Scale Factor 30,000 30TB Row Count
Catalog Returns	4,319,925,093
Catalog Sales	43,200,404,822
Inventory	1,627,857,000
Store Returns	8,639,952,111
Store Sales	86,399,341,874
Web Returns	2,160,007,345
Web Sales	21,600,036,511

Queries

The testing suite has 99 queries—4 of which have two parts (14, 23, 24, and 39). This brings a total of 103 queries. The queries used for the tests were compliant with the

[3] More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

standards set out by the TPC Benchmark™ DS (TPC-DS) specification[4] and included only minor query modifications as set out by section 4.2.3 of the TPC-DS specification document. For example, minor query modifications included vendor-specific syntax for date expressions. Also in the specification, some queries require row limits and, thus, vendor specific syntax was used (e.g., TOP, FIRST, LIMIT, and so forth) as allowed by section 4.2.4 of the TPC-DS specification.

Cluster Environments

Our benchmark included five (5) different cluster environments. The 3 non-anonymized ones are shown here. The cluster sizes were chosen to achieve similar hourly costs for each vendor, to the extent possible.

Table 2. Platform Summary

	Cloudera Data Platform	**Azure Synapse Analytics**	**Amazon Redshift**	**DW1**	**DW2**
Version Tested	**1.1.2-h2-b3**	**10.0.15391.0**	**1.0.20091**		
Cloud	**AWS**	**Azure**	**AWS**	**AWS**	**GCP**
Cluster Size	**EC2 r5d.4xlarge (64 nodes)**	**DW7500c**	**ra3.4xlarge (38 nodes)**		
Price Per Hour	**$123.26**	**$113.25**	**$123.88**	**$192.00**	**$120.00**

[4] The TPC Benchmark™ DS (TPC-DS) specification we used was found at http://tpc.org/tpc_documents_current_versions/pdf/tpc-ds_v2.13.0.pdf.

Test Results

This section analyzes the query results from the execution of the test queries (derived from the TPC-DS) described above. The primary metric used was the best aggregate total of each of the three runs. Three power runs were completed on each platform. Each of the 99 queries was executed three times in order (1, 2, 3) against each vendor cloud platform.

Table 3. Price-Performance Field Test Results for all Queries

Amounts shown represent the cost in USD for running each individual query. Red represents the most costly and green (often faint) is the least costly (best). The “Cl” column is Cloudera, “RS” is Redshift and “Sy” is Synapse.

Query	DW1	Cloudera	DW2	Redshift	Synapse	DW1	Cl	DW2	RS	Sy
1	0.216	0.164	0.420	0.148	0.399					
2	2.316	0.216	1.770	1.007	1.606					
3	0.287	0.078	0.224	0.546	0.181					
4	8.875	3.610	3.789	8.045	5.862					
5	0.684	0.606	4.674	0.343	1.007					
6	0.226	0.719	1.248	0.024	0.425					
7	0.471	0.173	0.876	0.275	0.599					
8	0.209	0.081	0.649	0.068	1.805					
9	4.991	1.619	1.762	7.219	1.875					
10	0.461	0.136	2.242	0.260	0.496					
11	5.216	2.463	2.310	4.185	4.058					
12	0.099	0.071	0.440	0.014	0.132					
13	0.809	0.444	1.254	0.630	0.607					
14	6.145	3.878	37.093	4.038	1.816					
14b	4.632	2.435	3.755	3.991	1.560					
15	0.268	0.745	0.584	0.213	0.416					
16	0.479	0.429	2.933	0.568	2.433					
17	1.182	0.554	0.916	0.170	0.874					
18	0.555	0.308	3.018	0.257	0.535					
19	0.263	0.139	0.435	0.058	0.318					
20	0.125	0.075	0.605	0.015	0.173					
21	0.059	0.046	0.175	0.006	0.098					
22	0.163	0.295	0.288	0.035	0.117					
23	7.596	6.801	6.163	7.367	3.990					
23b	7.567	6.943	6.092	7.762	3.423					
24	3.528	2.779	12.727	2.504	1.887					
24b	3.037	2.782	12.708	2.508	1.366					
25	0.842	0.439	1.219	0.215	0.641					
26	0.348	0.128	0.559	0.133	0.240					
27	0.605	0.177	2.297	0.253	0.467					
28	3.726	1.257	0.848	5.455	1.570					
29	1.036	0.755	2.312	0.274	0.947					
30	0.279	0.248	1.266	0.189	0.622					
31	0.884	0.925	1.579	0.605	1.559					
32	0.250	0.075	1.260	0.017	0.207					
33	0.529	0.225	2.303	0.085	0.505					
34	0.578	0.278	0.642	0.217	0.636					
35	0.859	0.363	2.976	0.561	0.987					
36	0.505	0.209	1.230	0.305	0.548					
37	0.193	0.209	0.427	0.201	0.138					
38	2.406	1.434	2.289	1.748	3.539					
39	0.205	0.345	0.336	0.013	0.145					
39b	0.172	0.345	0.327	0.010	0.162					
40	0.194	0.138	2.305	0.131	0.323					
41	0.028	0.025	0.094	0.006	0.040					
42	0.119	0.048	0.266	0.025	0.120					
43	0.564	0.127	0.440	0.259	0.329					
44	1.792	0.584	0.619	1.997	0.823					
45	0.297	0.187	2.312	0.146	0.444					
46	1.055	0.335	0.834	0.333	0.772					
47	1.283	1.486	1.257	1.885	1.655					
48	0.556	0.286	0.881	0.406	0.334					

Query	DW1	Cloudera	DW2	Redshift	Synapse	DW1	Cl	DW2	RS	Sy
49	1.742	0.288	0.877	0.299	0.969					
50	1.635	2.052	2.865	1.183	0.882					
51	0.523	0.343	5.637	0.464	0.523					
52	0.114	0.050	0.336	0.022	0.088					
53	0.257	0.116	0.342	0.131	0.141					
54	0.510	1.502	1.253	0.059	1.096					
55	0.126	0.049	0.343	0.019	0.130					
56	0.470	0.240	1.128	0.034	0.564					
57	0.774	0.912	0.890	1.221	1.417					
58	0.320	0.226	1.232	0.018	0.388					
59	3.002	0.288	2.982	1.553	4.259					
60	0.474	0.242	0.915	0.097	0.603					
61	0.545	0.239	0.916	0.098	0.417					
62	0.344	0.160	0.238	0.214	0.229					
63	0.211	0.121	0.360	0.135	0.149					
64	4.055	3.747	4.638	3.108	4.437					
65	1.375	1.263	0.804	0.404	0.489					
66	0.602	0.201	0.611	0.230	0.506					
67	4.701	7.541	6.411	6.026	20.355					
68	0.411	0.250	0.828	0.089	1.247					
69	0.473	0.208	1.736	0.549	0.577					
70	0.915	0.333	1.579	0.491	0.274					
71	0.478	0.176	0.822	0.147	0.524					
72	1.230	0.623	195.047	0.555	1.378					
73	0.232	0.177	0.628	0.054	0.508					
74	3.034	1.941	2.071	2.364	3.963					
75	3.499	2.002	3.376	1.980	3.440					
76	1.677	0.472	0.837	1.151	0.935					
77	0.453	0.180	2.303	0.070	0.528					
78	24.332	5.866	8.682	7.118	7.793					
79	1.195	0.430	1.248	1.087	1.370					
80	1.335	0.422	62.176	0.274	1.205					
81	0.263	0.392	0.893	0.278	0.833					
82	0.236	0.302	0.602	0.422	0.199					
83	0.142	0.179	0.286	0.017	0.272					
84	0.172	0.623	0.261	0.075	0.119					
85	0.611	0.237	0.604	0.302	0.493					
86	0.240	0.145	0.588	0.150	0.506					
87	2.059	0.533	2.136	1.554	3.834					
88	4.535	1.838	1.259	4.467	1.806					
89	0.316	0.133	0.454	0.186	0.179					
90	0.736	0.105	0.247	0.289	0.286					
91	0.190	0.115	0.280	0.012	0.147					
92	0.234	0.070	0.603	0.013	0.149					
93	2.048	2.964	3.744	1.208	0.897					
94	0.538	0.331	2.304	0.244	1.201					
95	7.406	0.428	21.774	0.242	1.283					
96	1.028	0.272	0.276	0.585	0.253					
97	1.284	0.694	2.942	0.556	0.885					
98	0.236	0.103	0.606	0.044	0.185					
99	0.712	0.303	0.461	0.534	0.516					

Price-Performance

To conduct comparative performance testing, we typically make the attempt to align the hardware and software as much as possible between the platforms. However, achieving a same like-for-like configuration with fully-managed cloud data warehouse platforms is very difficult. Thus, we aligned as closely as we could on price-per-hour as a basis for likeness. System cost can be a difficult aspect to compare systems, because vendor platforms vary on their pricing and licensing models. However, all platforms have a consistent on-demand hourly cloud pricing that we can use to determine price per performance.

Cloudera has a clear pricing model. However, there are both software costs (for the use of Cloudera Data Warehouse) and infrastructure costs (in our case, Amazon Web Services EC2 instances.) For the Cloudera software usage of the platform, there is a $0.72 per executor node (64 total) per hour. On the infrastructure side, there were 64 r5d.4xlarge executor nodes plus two (2) coordinator nodes of the same type—for a total of 66 at $1.152 per hour in US East (at the time of our testing). There are also three (3) m5.2xlarge management nodes that serve as orchestrators across three availability zones which cost $0.384 per hour. Altogether, it cost $123.26 per hour to run tests on a Cloudera Virtual Warehouse.

For Amazon Redshift, we simply paid a set dollar amount per hour by the instance class and node count that we configured. For example, considering the 38-node ra3.4xlarge configuration in the US East region (with rates at the time of the testing) we used for the 30TB test, we paid $3.26 per hour with 38 nodes, so $123.88 per hour.

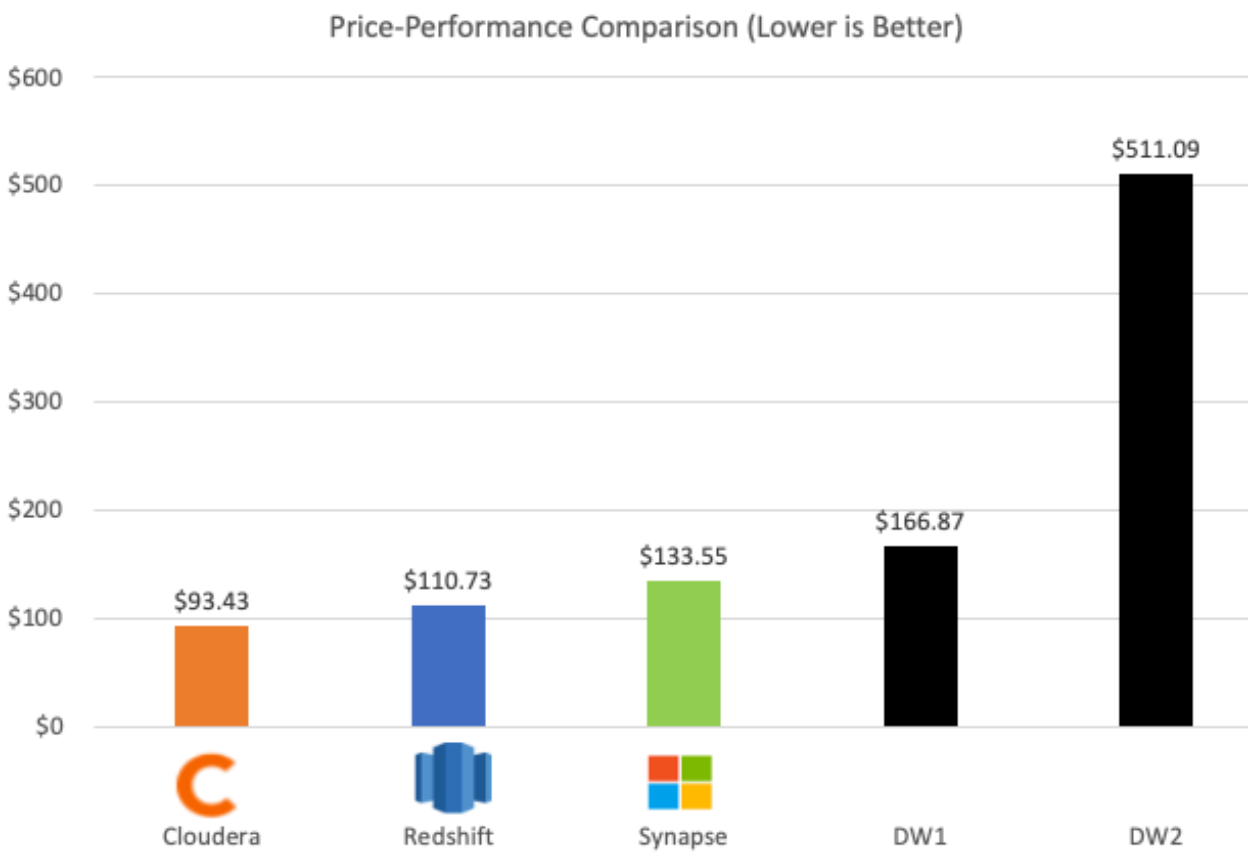
Azure Synapse Analytics charges $0.0151 per Data Warehouse Unit (DWU) per hour. Since we used a DW7500c with 7,500 units, we paid $113.25 in East US region.

With the hourly cost of the configuration, to calculate the price-per-performance, we used the following formula:

$$\frac{\text{Elapsed time of test (seconds)} \times \text{Cost of platform (\$/hour)}}{\text{3,600 (seconds/hour)}}$$

The elapsed time of the test is actually the duration of the fastest run of all 99 queries. The following tables detail the price-performance for the different tests.

Table 3. Price Performance


Price-Performance Comparison (Lower is Better)
$600
$500
$400
$300
$200
$100
$0
$93.43
$110.73
$133.55
$166.87
$511.09
Cloudera
Redshift
Synapse
DW1
DW2

Conclusion

Cloud data warehouses are a way for enterprises to achieve advanced analytics while avoiding large capital expenditures, provision quickly, and provide performance at scale for advanced analytic queries. Relational databases with analytic capabilities continue to support the advanced analytic workloads of the organization with performance, scale, and concurrency. In a representative set of corporate-complex queries from the well-known TPC-DS standard, Cloudera consistently performed equally, if not better than, the competition, and it proved to be the best value in terms of price per performance.

Overall, the test results were insightful in revealing query execution performance and some of the differentiators for Cloudera, Synapse, Amazon Redshift, and 2 anonymized public cloud data warehouses.

In terms of price per performance, Cloudera ran the Field Test 20% cheaper than the nearest competitor, Amazon Redshift, 40% cheaper than Synapse, 80% cheaper than one of the competitors and 5.5 times cheaper than the other.

Price and performance are critical points of interest when it comes to selecting an analytics platform, because they ultimately impact total cost of ownership, value, and user satisfaction. Our analysis reveals Cloudera to be very powerful and comparative in value.

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.